Un sistema di caching WWW nazionale

P. Tentoni
CILEA, Segrate

Abstract
E' stato costituito a livello nazionale un gruppo di lavoro denominato GARR-CACHE per il coordinamento delle iniziative di caching WWW nell'ambito della rete della ricerca scientifica. Il CILEA partecipa a questo gruppo di lavoro e sta operando per la costituzione di un servizio di wwwcache CILEA, inserito nella gerarchia nazionale ed internazionale, aperto a tutta la rete GARR, e promuove l'uso della cache per arrivare all'obiettivo di una maggior soddisfazione per l'utente Internet, nonché ad un risparmio considerevole di banda trasmissiva ed altre risorse di rete.

Perché nascono le cache WWW

E' ormai da tempo che i Web Browser (Netscape, MS Internet Explorer, Mosaic) sono diventati parte ed essenza stessa del lavoro di ricerca e di studio per un numero sempre maggiore di utenti, e non solo scientifici.
Studi accurati hanno accertato inoltre che la maggior parte del traffico di rete è proprio generato da questi tipi di servizi, ed i tempi di risposta del server remoto e della rete sono divenuti particolarmente importanti e cruciali.

Il fenomeno è accentuato proprio dalla stessa architettura Client-Server dei servizi di rete, che prevede di realizzare connessioni dirette tra ogni client ed il server. Questo *provoca di conseguenza, sia la trasmissione delle stesse informazioni a più client, spesso molto vicini, che la richiedano contemporaneamente* (di qui il consumo di banda trasmissiva), *sia il sovraccarico del server ricevente* che in questo modo tanto più è popolare, quanto più deve soddisfare un numero elevato di richieste al secondo.

Questo tipo di problemi è evidentemente indipendente dal singolo protocollo (ftp, http, o altro), ma le soluzioni adottabili possono essere differenti nei diversi casi, e così pure diverso è l'impatto sull'utente finale del ritardo indotto dall'eccessivo carico e consumo di risorse di rete.

Infatti se l'utente è disposto ad aspettare pazientemente un mail o un file transfer, lo è molto meno nell'attesa per la visualizzazione di una pagina HTML, per la quale l'interazione è evidentemente molto maggiore e di conseguenza assai maggiore la frustrazione di chi attende il suo riempimento progressivo davanti al proprio video.

Queste motivazioni hanno già da alcuni anni spinto diverse comunità di utenti in Internet (e le prime sono state non a caso quelle che avevano connessioni alla rete con capacità di banda molto limitate) a *realizzare sistemi di caching WWW specializzati ed aventi lo scopo di minimizzare in traffico generato da più client della comunità che richiedono pagine comuni in tempi vicini.*

Il protocollo HTTP inoltre possiede già implicito il meccanismo del proxy, ed è quindi stato da subito più facile realizzare un sistema di caching piuttosto che non la replicazione dei server (come può essere un sito di mirroring FTP).

Come si è già accennato, sono nati a livello internazionale numerosi progetti sia in ambito nazionale sia sovranazionale per il coordinamento e la razionalizzazione di tali servizi, che da una cooperazione e scambio reciproco di informazioni possono trarre particolare vantag-

gio. Alcuni dei più rilevanti che possiamo citare sono:

- NLANR, nell'ambito di NFSnet (U.S.): **www.nlanr.net/Cache**
- CHOC, nell'ambito di TERENA (Europa) **www.terena.nl/tech/projects/chocs/workplan.html**
- HENSA, per la rete universitaria Inglese (U.K**.) www.hensa.ac.uk/wwwcache/intro.html**
- RENATER (Francia): **cache.cnrs.fr/**

Visto l'interesse generale della comunità scientifica a realizzare un sistema coordinato di cache per il miglioramento delle performance generali della rete e per la riduzione del traffico indotto sulle costose tratte internazionali, il CILEA ha deciso di offrire il suo contributo riservando allo scopo sia le risorse umane sia gli apparati hardware per la creazione di una cache di livello nazionale.

Lo scopo del progetto GARR CACHE sarà quello ***di realizzare un sistema di root cache nazionali*** (cinque, per il momento, opportunamente distribuiti sul territorio nazionale) ed inoltre ***di promuovere la diffusione a livello locale di cache di secondo o terzo livello che interagiscano con le cache primarie creando un sistema coordinato di cache WWW,* avente come fine il mantenimento delle informazioni più richieste a livello locale distribuito,** in modo da ridurre il consumo di banda sulle costose tratte internazionali, nonché l'eccessivo carico sui server remoti più popolari.

Il meccanismo

Il servizio che il CILEA ha realizzato può agire come "proxy" rispetto ai client Web propri e di tutti gli enti che non si sono ancora dotati di un sistema autonomo e locale di caching.
Questo fa sì che il singolo client, invece di richiedere direttamente al server finale il documento Web, si rivolga invece al suo proxy server, ovvero ad esempio alla cache CILEA (**wwwcache.cilea.it**).
E' il cache-server che al posto suo andrà a richiedere al sito remoto la pagina http (o il servizio ftp, gopher), consegnandola al client che l'ha richiesta, ma anche mantenendola nella propria memoria (disco e/o RAM), per poter soddisfare rapidamente (e questa volta *senza causare traffico dalla sorgente al client*) successive richieste per quella URL.

Un server cache WWW oltre ad agire come proxy per i propri client, è inserito in una struttura gerarchica opportunamente coordinata.
Questo significa che si pone in relazione con altri server cache della stessa specie (utilizzando lo speciale protocollo ICP), ovvero può, se autorizzato, sfruttarne le risorse, cioè richiedere pagine presenti nella altrui cache, moltiplicando così lo spazio disco virtualmente disponibile.
Un server può agire come "***parent***" (genitore) per altri server, può agire come "***sibling***" (fratello/sorella) restandone allo stesso livello, o viceversa può agire come figlio rispetto ad altri server.
La differenza tra questi diversi comportamenti è la seguente: il padre ha sempre il dovere di fornire il documento richiestogli dai figli, sia che lo possegga già nella propria cache (HIT) sia che lo debba richiedere alla fonte (MISS), conservandolo poi per un successivo potenziale utilizzo.
Un fratello invece ha il solo dovere di soddisfare richieste di documenti già presenti sulla propria cache (HIT) senza intraprendere nessuna altra azione in caso contrario.

E' stato osservato che è utile ed opportuno strutturare le cache nazionali in modo che esistano in generale due ed al massimo tre livelli:

1. i root cache (più di uno possibilmente per bilanciare il carico e per backup) avente il compito eventualmente di colloquiare con root cache internazionali
2. un livello di cache regionale
3. un livello di cache nelle diverse organizzazioni di una certa importanza (quale ad es. una grande Università che abbia aperto l'utilizzo di Internet ai propri studenti e che si aspetta di generare un traffico omogeneo e molto localizzato)

I requisiti

La localizzazione di una cache di primo livello deve soddisfare sia requisiti ovvi relativi alla connettività interna al GARR ed anche esterna (vicinanza alle uscite internazionali GARR), sia considerazioni di tipo geografico.
Inoltre la configurazione hardware relativamente alla memoria centrale e memoria di massa deve essere adeguata al suo livello parti-

colarmente alto nella gerarchia, che implica un elevato numero di richieste da soddisfare (quelle di tutta la discendenza).

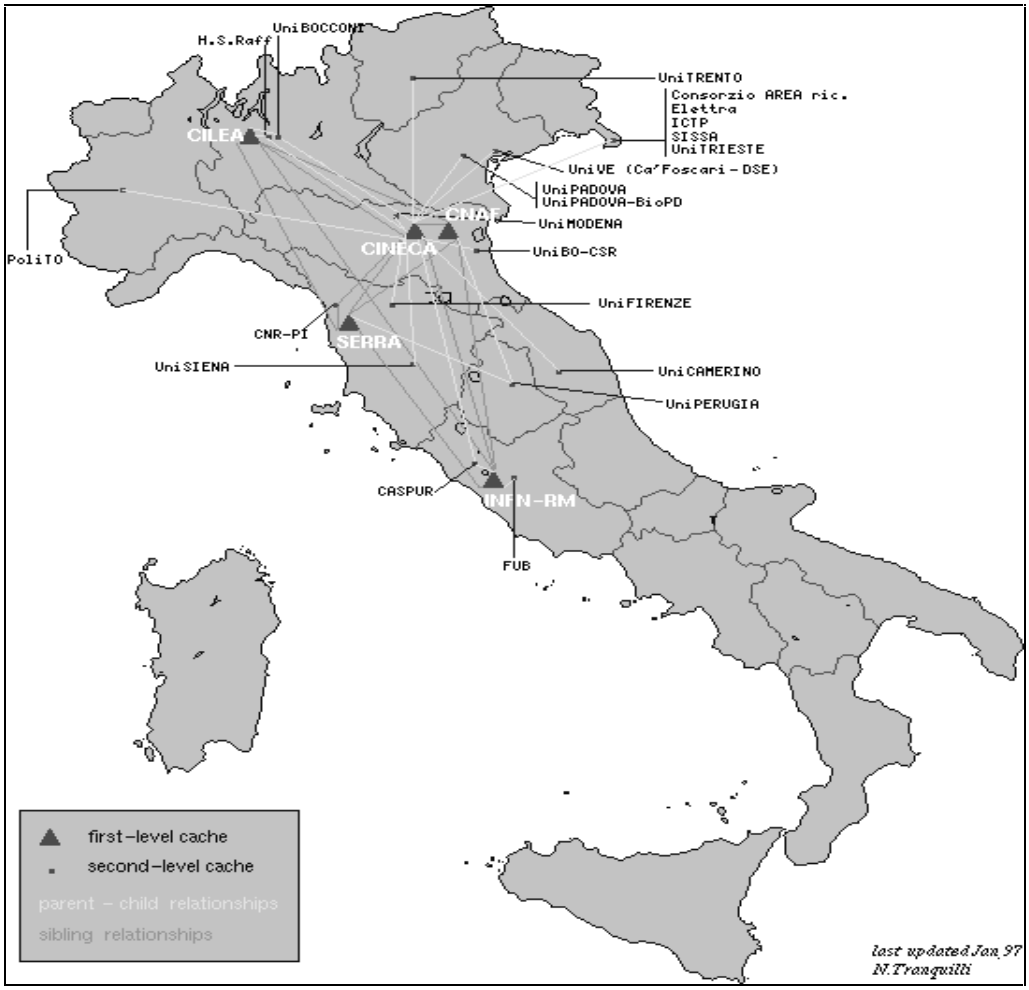
La soluzione attualmente adottata dal GARR prevede la presenza di cinque cache Nazionali (root cache o cache di primo livello) così dislocate:

- Milano - CILEA (wwwcache.cilea.it)
- Bologna - CINECA (proxy.cineca.it)
- Bologna - CNAF (wwwcache.infn.it)
- Pisa - Università (proxy.unipi.it)
- Roma - INFN (wwwcache.roma1.infn.it)

cache nazionali e regionali i documenti html più richiesti.
Per la realizzazione di questo servizio il CILEA ha **dedicato** una propria stazione HP-735 con le seguenti caratteristiche hardware:

- 128Mbyte di RAM
- 10 Gbyte di dischi
- una interfaccia di rete FDDI su cui è connesso il router di frontiera

Su di essa è stato installato e configurato il software specializzato ***Squid versione 1.1.0.***



Dato il ritardo in generale dell'Italia (e della rete della ricerca in particolare) **si ritiene particolarmente urgente ed importante dare la necessaria priorità a questo servizio nascente** che ad un costo iniziale (hardware e di risorse umane dedicate al servizio) dovrebbe far seguire, secondo gli auspici di tutti gli aderenti al progetto, una effettiva riduzione del carico sulle linee internazionali, localizzando sulle

Perché aderire all'iniziativa di caching

Sul perché aderire ci siamo già dilungati nel descrivere le motivazioni che hanno spinto un po' tutti i gestori di reti a realizzare questo servizio, taluni mossi dalla urgente necessità di ovviare alla indisponibilità di risorse trasmissive adeguate.

La soddisfazione del singolo utente che utilizza una cache ben configurata è pure un buon indice dell'importanza di tale servizio.

Le frustranti attese rispetto ai tempi di risposta standard della rete, soprattutto nella fascia oraria pomeridiana, quando ormai è quasi impossibile lavorare sulla rete, si possono eliminare tanto più popolare è la richiesta che si è fatta, avendo una elevata probabilità di vederle soddisfatte dalla cache e quindi di averle recapitate in tempi rapidissimi.

Cosa deve fare il navigatore Web

Ottenere i benefici di una cache vicina è facile, e **conviene sempre** perché se il documento è in cache (o nelle cache parenti) questo vi viene **subito consegnato**, senza attendere i tempi di risposta dei link internazionali.

Se invece non è in cache, la cache ve lo procurerà come un qualunque altro proxy server, andando direttamente alla fonte come farebbe il vostro Browser ma poi tenendolo memorizzato localmente, per un suo successivo utilizzo.

E' facile, dicevamo, si deve solo riconfigurare il proprio Browser in modo che indichi come proxy il seguente indirizzo:

Proxy http: **wwwcache.cilea.it** Port: **8080**
Proxy ftp: **wwwcache.cilea.it** Port: **8080**
Proxy Gopher: **wwwcache.cilea.it** Port: **8080**

Se è possibile farlo, indicate poi tra i domini *No Proxy* il vostro dominio, al quale accederete sempre più velocemente che attraverso la cache.

Sulle più recenti versioni di **Netscape** è possibile *attivare automaticamente la configurazione del proxy:*

1. Options
2. Network Preferences:
3. Proxy

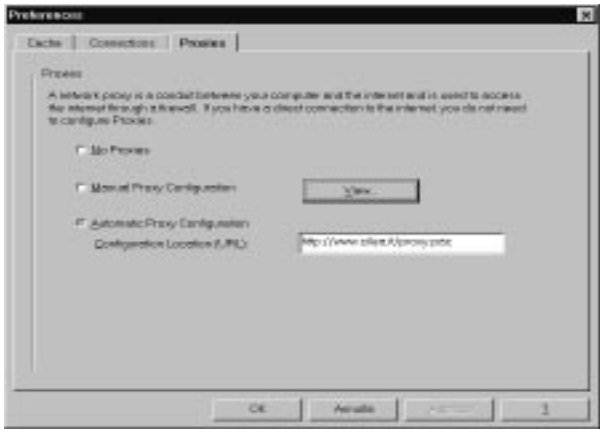
Selezionare il bottone:

Automatic Proxy Configuration

e dichiarare sulla stessa riga, come URL del configuratore il seguente:

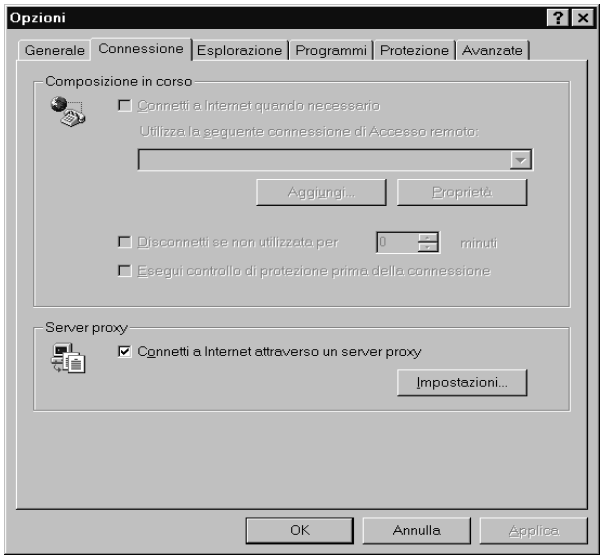
http://www.cilea.it/proxy.pac

Attualmente esiste solo per Netscape la possibilità di utilizzare questo sistema automatico di configurazione del browser.

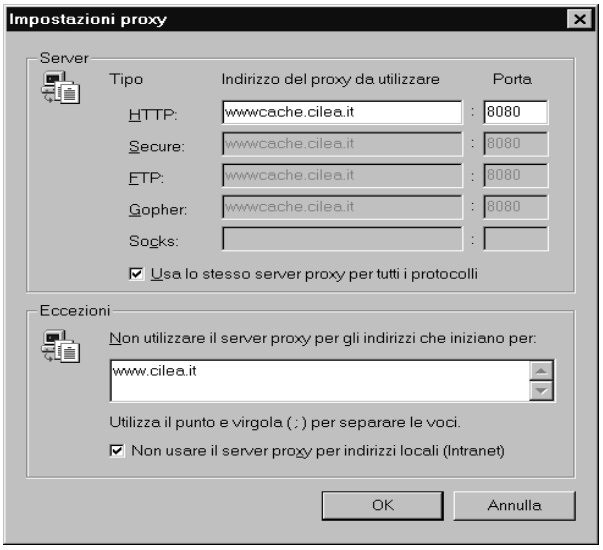


Si tratta di un piccolo Java script che consente anche di fare, volendo, load-balancing su eventuali cache multiple, e garantisce comunque la disponibilità almeno diretta del documento in caso di *mancato funzionamento del server cache e per la successiva mezzora.*

Per chi utilizza invece **Microsoft Internet Explorer** (versione 3.0) occorre:
1) selezionare dalla barra Menu: *Visualizza*
2) selezionare quindi: *Opzioni....*
3) selezionare la scheda: *Connessione*



Su questa va attivato il quadratino *Connetti ad Internet attraverso un server proxy* per poi andare in *Impostazioni....*



Qui occorre indicare come HTTP proxy server wwwcache.cilea.it (PORT 8080).
Poi è sufficiente barrare la casella *Usa lo stesso server proxy per tutti i protocolli.*

Come partecipare al progetto di caching nazionale

Come si è già accennato può essere efficace che anche i gestori di grandi reti locali, quali sono tipicamente le reti d'Ateneo, *si impegnino a creare un proprio sistema locale di caching WWW.*

Le Università infatti sono sempre più aperte all'utilizzo di Internet da parte dei proprio studenti e possono generare un elevato traffico verso Internet, spesso largamente omogeneo.

Trarrebbero quindi da un caching locale delle pagine più richieste il doppio vantaggio di una più rapida risposta al Client richiedente la pagina e la diminuzione del traffico verso la rete globale per tutte quelle transazioni che invece non sono normalmente memorizzabili su un sistema di cache (quali le ricerche o gli accessi richiedenti login), ovvero liberando banda per altre applicazioni.

Allo stato attuale hanno già realizzato sistemi di caching locale collocati nella gerarchia GARR nazionale l'Università Bocconi, anche se in fase sperimentale, e l'Ospedale S.Raffaele.
Per tutti i gestori di cache locali, o per coloro che hanno intenzione di realizzarla a breve, purché appartenenti alla rete GARR è possibile configurarsi come figli della cache CILEA, se questo ha un senso dal punto di vista della connettività.
Ed è comunque possibile richiedere un *supporto tecnico a gruppo di lavoro GARR per l'attuazione veloce del servizio.*
In questo caso è bene che chi realizza una cache di secondo livello registri la propria attività attraverso il modulo informativo che trova al seguente indirizzo:

wwwcache.lnf.infn.it/garrcache/cgi-bin/rac.pl

In questo modo sarà evidenziata sulla mappa della gerarchia italiana la sua collocazione.
Tale mappa è visualizzabile all'indirizzo:

www.cineca.it/proxy/garr/topology.html

Nel caso si voglia realizzare una relazione di parentela con la cache CILEA (ponendosi come figli) i parametri di configurazione devono indicare:

- HostName: **wwwcache.cilea.it**
- HTTP port**: 8080**
- ICP port: **3130**

Se si sta utilizzando Squid o Harvest ciò equivale alle seguenti righe in configurazione:

cache_host wwwcache.cilea.it parent 8080 3130
cache_host_domain wwwcache.cilea.it

In tal caso la relazione è immediatamente operativa, senza bisogno di alcuna richiesta formale al CILEA, poiché le access list del cache server prevedono l'autorizzazione alle richieste ICP provenienti da qualunque indirizzo della rete GARR.

Ulteriori informazioni sul servizio di cache CILEA, possono essere reperite alla pagina:

http://www.cilea.it/servizi/a/proxy/

oppure rivolgendosi ai gestori della cache CILEA, raggiungibili attraverso l'indirizzo di posta elettronica:

noc@cilea.it